Infinite Latent Process Decomposition

Tomonari Masada, Yuichiro Shibata, and Kiyoshi Oguri
Department of Computer and Information Sciences
Nagasaki University
Nagasaki, Japan
{*masada,shibata,oguri*}*@cis.nagasaki-u.ac.jp*

***Abstract*—This paper presents *infinite latent process decomposition (iLPD)*, a new microarray analysis method, as an extension of latent process decomposition [1]. Our method assumes an infinite number of latent processes. Further, our new collapsed variational Bayesian inference improves the inference proposed in [2] in the treatment of Dirichlet hyperparameters. We also give the results of the comparison experiment.**

***Keywords*-microarray data; Bayesian nonparametrics;**

I. Introduction

In this paper, we propose a new Bayesian microarray analysis as an extension of latent process decomposition (LPD) [1] by assuming an infinite number of latent processes. LPD can be understood as a "bioinformatics variant" of LDA [3]. Therefore, our proposal extends LPD just as HDP [4] extends LDA. We call our method *infinite latent process decomposition (iLPD)*. Further, we devise a new inference based on CVB for HDP [5], some of whose update formulas can be imported as is. Our CVB can be applied to both iLPD and LPD and improves the CVB for LPD [2] in two aspects. First, we introduce auxiliary variables as in [5] for Dirichlet hyperparameters. Second, we approximate the variational lower bound in a more natural manner. The approximation proposed in [2] depends on the ordering of genes. We remove this dependence by using an efficient second-order approximation. These two aspects are independent of the assumption of an infinite number of latent processes. The model details are given in the extended version [6].

II. Model

This section describes iLPD in a generative manner.

For each sample d, a probability distribution defined over infinite latent processes is drawn from the Dirichlet process $\texttt{DP}(\alpha, \pi)$. Each latent process corresponds to a hidden cluster of genes. We adopt the truncated posterior representation and set the number of latent processes to the finite number K as in [5]. Therefore, the drawn distribution can be regarded as a multinomial distribution $\texttt{Multi}(\theta_d)$ defined over K latent processes. For the center measure π, we use the stick-breaking representation [7]: $\pi_k = \tilde{\pi}_k \prod_{l=1}^{k-1}(1-\tilde{\pi}_l)$ and $\tilde{\pi}_k \sim \texttt{Beta}(1,\gamma)$ for each k. The parameter γ of $\texttt{Beta}(1,\gamma)$ is in turn drawn from the Gamma distribution $\texttt{Gamma}(a_\gamma, b_\gamma)$. The concentration parameter α of $\texttt{DP}(\alpha,\pi)$ is drawn from the Gamma distribution $\texttt{Gamma}(a_\alpha, b_\alpha)$.

The observed expression data are generated as follows. For each pair of gene g and latent process k, the mean and the precision of the Gaussian distribution $\texttt{Gauss}(\mu_{gk}, \lambda_{gk})$ are drawm from the Gaussian prior $\texttt{Gauss}(\mu_0, \rho)$ and the Gamma prior $\texttt{Gamma}(a_0, b_0)$, respectively. The precision ρ of the Gaussian $\texttt{Gauss}(\mu_0, \rho)$ is in turn drawn from the Gamma distribution $\texttt{Gamma}(a_\rho, b_\rho)$. Then, for each pair of sample d and gene g, a latent process z_{dg} is drawn from the multinomial $\texttt{Multi}(\theta_d)$. Based on this draw, a real number x_{dg} is drawn from the Gaussian $\texttt{Gauss}(\mu_{gz_{dg}}, \lambda_{gz_{dg}})$. x_{dg} corresponds to the expression data in the microarray.

With respect to the CVB inference for iLPD, the details are referred to the extended version of this paper [6].

III. Comparison Experiment

Table I gives the results obtained in our experiment comparing iLPD with LPD proposed in [2]. We used the 11 datasets available at http://www.gems-system.org/. The results in Table I are also given in Figure 1 by charts.

In our experiment, the expression data are normalized to the Gaussian distribution of zero mean and unit variance for each gene. For comparing the efficiency of iLPD with that of LPD, we randomly select 10% expression data as test data from each microarray and use the rest 90% as training data. After conducting a CVB inference starting from a random initialization of z_{dg}s, we evaluate the density function of the posterior at each of the test data and calculate the geometric mean of these density evaluations over all test data. This geometric mean is our comparison measure, which is devised as a counterpart of *perplexity*, a measure often used in natural language processing. A larger geometric mean corresponds to a better generalization power. For K, we tried the following three settings: 10, 20, and 40.

The inference is implemented in C for execution-time efficiency and is further parallelized with OpenMP library on Intel Core i7 920 CPU. For example, *14 Tumors*, the largest dataset among those used in our experiment, requires around 6,800 seconds for 150 iterations of our CVB for $K = 40$.

We ran the inference 25 times each starting from a different random initialization. The mean and the standard deviation of the 25 corresponding evaluations are presented in Table I and Figure 1. As is shown in Table I and Figure 1, iLPD can adapt to all settings of K. In contrast,

Table I
COMPARISON EXPERIMENT RESULTS

11 Tumors	$K = 10$	$K = 20$	$K = 40$
iLPD	0.3119±0.0018	0.3143±0.0009	0.3143±0.0011
LPD	0.3134±0.0011	0.3128±0.0009	0.3074±0.0011
14 Tumors	$K = 10$	$K = 20$	$K = 40$
iLPD	0.4792±0.0022	0.4910±0.0030	0.4954±0.0025
LPD	0.4786±0.0027	0.4909±0.0026	0.4860±0.0028
9 Tumors	$K = 10$	$K = 20$	$K = 40$
iLPD	0.1696±0.0152	0.1658±0.0155	0.1669±0.0139
LPD	0.1713±0.0103	0.1687±0.0098	0.1543±0.0122
Brain 1	$K = 10$	$K = 20$	$K = 40$
iLPD	0.2891±0.0032	0.2879±0.0048	0.2884±0.0054
LPD	0.2898±0.0027	0.2852±0.0032	0.2769±0.0024
Brain 2	$K = 10$	$K = 20$	$K = 40$
iLPD	0.2450±0.0038	0.2428±0.0080	0.2441±0.0064
LPD	0.2463±0.0040	0.2415±0.0079	0.2378±0.0063
Leuk. 1	$K = 10$	$K = 20$	$K = 40$
iLPD	0.2496±0.0049	0.2490±0.0066	0.2524±0.0040
LPD	0.2485±0.0027	0.2441±0.0037	0.2377±0.0021
Leuk. 2	$K = 10$	$K = 20$	$K = 40$
iLPD	0.3281±0.0025	0.3217±0.0026	0.3137±0.0031
LPD	0.3281±0.0026	0.3216±0.0024	0.3083±0.0035
Lung Can.	$K = 10$	$K = 20$	$K = 40$
iLPD	0.3511±0.0012	0.3519±0.0015	0.3528±0.0016
LPD	0.3522±0.0010	0.3504±0.0013	0.3452±0.0015
SRBCT	$K = 10$	$K = 20$	$K = 40$
iLPD	0.2670±0.0041	0.2680±0.0041	0.2682±0.0032
LPD	0.2690±0.0044	0.2586±0.0030	0.2354±0.0034
Prostate	$K = 10$	$K = 20$	$K = 40$
iLPD	0.4531±0.0113	0.4537±0.0150	0.4462±0.0091
LPD	0.4583±0.0101	0.4492±0.0127	0.4420±0.0135
DLBCL	$K = 10$	$K = 20$	$K = 40$
iLPD	0.2704±0.0049	0.2709±0.0039	0.2729±0.0045
LPD	0.2711±0.0033	0.2674±0.0033	0.2576±0.0033

the evaluation of LPD drops for $K = 40$ by a large margin. That is, iLPD is unlikely to be affected by the setting of K.

IV. CONCLUSIONS

This paper provides a new Bayesian nonparametric model for the microarray analysis. The experimental results reveal that iLPD is less affected by the setting of K, i.e., the number of latent processes, when compared with LPD [2].

We are now engaged in improving the experiment settings, especially the dataset selection and the comparison strategy.

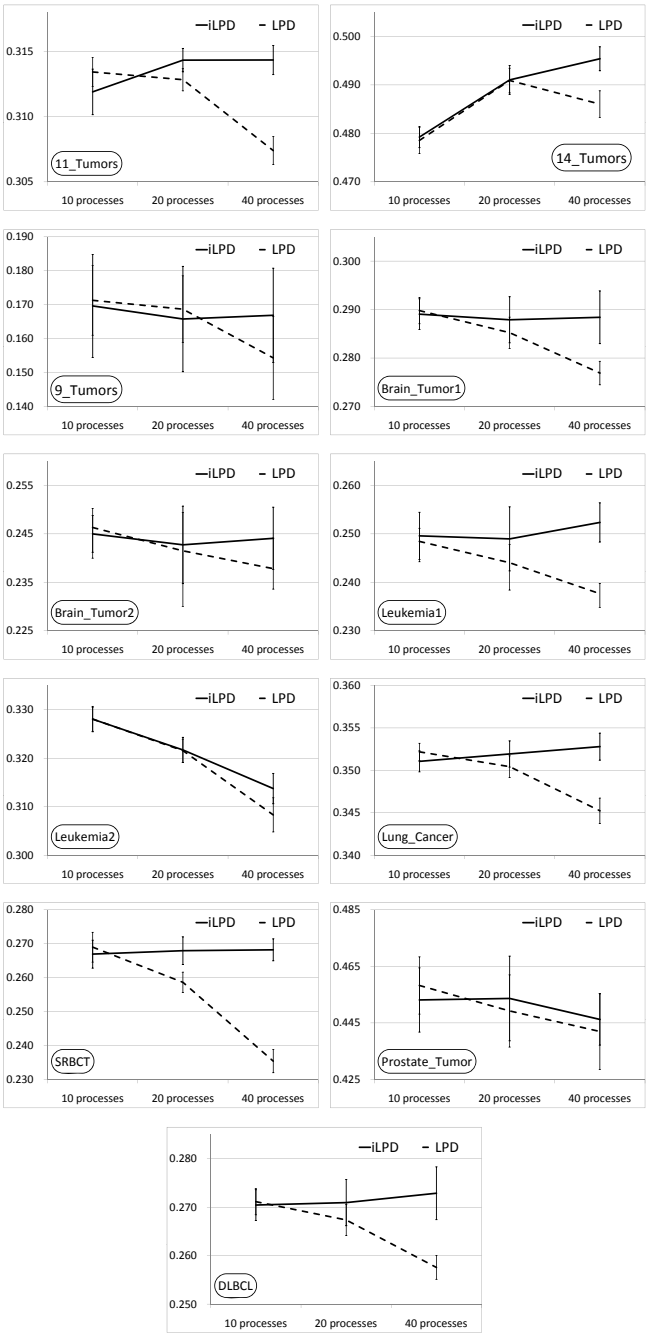

Figure 1. Comparison Experiment Results

REFERENCES

[1] S. Rogers, M. Girolami, C. Campbell, and R. Breitling, "The latent process decomposition of cDNA microarray data sets," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 2, no. 2, pp. 143–156, 2005.

[2] Y.-M. Ying, P. Li, and C. Campbell, "A marginalized variational Bayesian approach to the analysis of array data," *BMC Proceedings*, vol. 2(Suppl 4):S7, 2008.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *JMLR*, vol. 3, pp. 993 – 1022, 2003.

[4] Y.-W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *J. Amer. Stat. Assoc.*, vol. 101, no. 476, pp. 1566–1581, 2006.

[5] Y.-W. Teh, K. Kurihara, and M. Welling, "Collapsed variational inference for HDP," in *NIPS 20*, 2008, pp. 1481–1488.

[6] T. Masada, "Infinite latent process decomposition," 2010. http://www.cis.nagasaki-u.ac.jp/masada/CVBiLPD.pdf

[7] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.